ChinaXiv合作期刊

知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

数字人文领域的知识图谱：研究进展与未来趋势

朱丽雅[1] 张珺[1] 洪亮[1] 罗绍辉[2] 兰度[2]

1. 武汉大学信息管理学院 武汉 430072
2. 南宁市勘察测绘地理信息院 南宁 530022

摘要：[目的/意义] 对数字人文领域的知识图谱研究进行系统性回顾，旨在提供未来可能的研究方向和开放的研究主题。[方法/过程] 以国内外会议、期刊发表的相关文献为研究对象，采用综合归纳法，系统梳理数字人文领域知识图谱的理论与实践发展。阐述数字人文领域知识图谱的相关概念，并根据当前的研究热点，从数据资源建设、关键构建技术、平台智能应用 3 个方面揭示其研究动向，并对未来研究趋势进行展望。[结果/结论] 总结数字人文知识图谱研究的未来发展趋势，即未来将呈现出多源数据集成、多模态知识融合、多学科交叉应用的发展趋势。

关键词：数字人文　知识图谱　智慧数据　数据资源建设　语义挖掘

分类号：G252.8

引用格式：朱丽雅, 张珺, 洪亮, 等. 数字人文领域的知识图谱：研究进展与未来趋势 [J/OL]. 知识管理论坛, 2022, 7(1): 87-100[引用日期]. http://www.kmf.ac.cn/p/277/.

1 引言

数字人文（Digital Humanities, DH）起源于 20 世纪 40 年代末的人文计算。人文计算侧重于对计算与人文学科之间的交叉领域进行研究、学习与创新[1]。随着时代的信息化程度不断加深，以及数字资源的不断增加，仅凭人文计算难以完成更高层次的学术发现。因此，数字人文的概念应运而生，它是在计算机技术、网络技术、多媒体技术等新兴技术支撑下开展人文研究而形成的新型跨学科研究领域[2]。在我国，如何通过数字化激发创新创造活力，推动文化产业

基金项目：本文系 2020 年国家档案局科技项目“基于时空数据的智慧城市档案知识图谱构建及应用服务体系研究”（项目编号：2020-X-053）、湖北省重点研发计划项目“文旅科技大数据关键技术研发与应用示范”（项目编号：2020BAB117）和南宁市科学研究与技术开发计划项目科技重大专项“基于 GIS 和 BIM 技术的城建大数据平台研究”（项目编号：20193010）研究成果之一。

作者简介：朱丽雅，硕士研究生；张珺，硕士研究生；洪亮，教授，博士，博士生导师，通信作者，E-mail: hong@whu.edu.cn；罗绍辉，高级工程师；兰度，高级工程师。

收稿日期：2021-09-30　　发表日期：2022-02-23　　本文责任编辑：刘远颖

迈向高质量发展，从而更好地满足人民群众日益增长的精神文化需求，成为一项重要课题。例如，2019 年中华人民共和国文化部发布的《文化部“十三五”时期文化产业发展规划》中强调要促进数字文化产业创新发展，包括推进“文化 +”和“互联网 +”战略，促进互联网等高新科技在文化产业各环节的应用。2020 年，国家“十四五”规划提出实施文化产业数字化战略。随着“数智时代”的到来和数字人文的兴起，数字人文研究中的数据基础设施和数字学术环境已经成为数字人文资源开发利用的重要方面。

在研究数字人文的过程中，结合知识图谱能为其带来新的方法与新的思考。一方面，知识图谱作为人工智能时代一种先进的知识组织方式，能够为数字人文研究提供优良的技术支持，去发掘那些以往在文本资源中看不见的模式和联系。另一方面，知识图谱作为智慧数据的表现形式，为数字资源的挖掘分析提供了基础，进行大规模的知识图谱构建能够提高建设智慧化数字人文系统的效率，并为该领域研究者以及其他想要了解人文学科的人员提供专业的、智能的知识服务。然而，数字人文领域知识图谱的研究成果虽然多，但比较分散，缺少一个系统的体系。因此，本文将深入开展数字人文领域知识图谱研究，并整合相关研究成果。

2 数字人文领域知识图谱概念辨析与文献收集

2.1 概念辨析

在图书馆和数字人文领域，知识图谱的概念深深植根于知识组织系统 [3]。数字人文领域知识图谱旨在利用知识图谱这一先进的知识组织方式，对原本分散的、异构的海量数据进行整合，从而满足领域学者的研究需求，并实现智能知识服务。与通用知识图谱相比，数字人文领域的知识图谱具有以下特点：

首先，在数据方面，研究者已经认识到了传统资源利用与开发模式的局限性，开始有意识地将数字人文领域普通的数字化资源转为智慧化资源。从以往只具有检索功能的数据库形式逐渐转变为具有推理分析功能的智能平台形式，充分利用新的信息技术来深入挖掘知识。

其次，数字人文领域知识图谱立足于学者导向的研究需求，其目的和通用知识图谱不同，不是要求涵盖各范围广泛的知识以实现全方面的知识检索，而是在实现大范围的知识覆盖的基础上，构建更为全面的知识体系，来搭建支持智慧化的领域知识服务平台。

最后，数字人文知识图谱所涉及的领域较为广泛，在构建知识图谱的过程中，需要充分考虑不同研究领域的影响。例如，周莉娜等 [4] 在构建唐诗知识图谱时提出，由于唐诗知识涉及到诗学、文献学、史学这三大领域，通过分析三大领域现存的未决问题，就能够较为全面地发掘出唐诗知识图谱的构建需求。因此，数字人文领域知识图谱与通用知识图谱在构建方法上也存在诸多不同，尤其体现在本体构建、知识抽取、知识推理等构建技术中。

2.2 文献收集

2.2.1 文献来源

（1）检索范围。本文的研究文献主要通过国内外数据库获取。考虑到研究的新颖性，选取了 2010 年至 2021 年的文献。国内文献来源于中国知网，选择图书情报类的学术核心期刊，如《中国图书馆学报》《情报学报》《数据分析与知识发现》等期刊；国外文献来源于 WOS、Elsevier、EBSCO 及 Springer 等数据库，选择 Information Science & Library Science 领域的学术核心期刊，如 *MIS Quarterly*、*Journal of Information Technology*、*International Journal of Information Management* 等期刊。

（2）检索关键词。国内数据库以“数字人文”“知识图谱”为检索词，国外数据库以“digital humanities”“knowledge graph”为检索词，分别采用标题、主题途径进行检索，并对检索结果进行筛选、去重、勘误，去除了与主题关联度较低的文献。考虑到仅采用以上两个关键词进行检索具有局限性，无法深入反映知识图谱

知识管理论坛，2022（1）:87-100
DOI: 10.13266/j.issn.2095-5472.2022.008

在数字人文领域中的具体研究内容，又选取“智慧数据”（smart data）、“本体”（ontology）、“知识抽取”（knowledge extraction）、“关联数据”（linked data）”等作为检索词来挖掘知识图谱在数字人文研究中的具体应用，保证检索结果可以较为全面地覆盖数字人文领域的代表性研究成果，并再次对检索结果进行筛选、去重、勘误。最终得到国内文献 131 篇、国外文献 187 篇作为初始样本。

2.2.2 研究热点简述

整体而言，数字人文领域知识图谱的研究呈现出多学科、文理交融的特点，涵盖了历史学、文献学、计算机科学、管理学、图书馆学等多种学科。它将过去研究中容易割裂的技术与文化进行了有机融合，利用其他学科丰富的数据资源与成熟的实践体系，为数字人文领域知识图谱研究带来有力的基础支撑，极大地丰富了该领域的研究内容，对推进数字人文智慧化研究体系具有重大意义。研究的主要热点集中在以下 3 个方面：

（1）数字人文领域数据资源建设。此类研究是国内外数字人文领域知识图谱的研究起点，主要探索与数字人文领域相关的各类数据资源建设，包括古籍文献、图像、视频、音频等各类结构化、半结构化及非结构化数据源。F. Kaplan[5] 将数字人文的大数据研究作为一个结构化的研究领域，提出了三个同心研究领域的划分。在其基础上，国内外学者就数字人文领域数据资源分类、特色、数字化方法等问题进行了深入研究，如董政娥等 [6] 针对数字人文特点，对数字人文文献资源进行了调查。数据资源建设作为数字人文知识图谱构建的基础步骤，能够为其提供数据源支持。

（2）数字人文知识图谱构建技术。此类研究是数字人文领域知识图谱研究中的重点，利用各类数字人文领域数据源，面向数字人文领域数据的特点，研究本体构建、知识抽取、消歧等问题，解决不同知识图谱的融合和跨语言实体的对齐问题。在这类文献中，国内的起步虽然较晚，但是针对我国的文化特色开创了不少针对性研究，如陈涛等 [7] 构建的 SinoPedia 平台，采用 RDF 三元组对目前公共领域的百科概念术语赋予唯一的 URI 进行资源的持久化，有助于中文知识图谱和中文领域本体的标准化和推广应用。

（3）数字人文知识图谱平台智能应用。此类研究是数字人文领域知识图谱研究发展的必然路径，主要着重于数字人文中的关联数据技术运用，以支持大规模、可重用的数字人文研究，如 R. Hoekstra 等 [8] 介绍了数字人文数据管理项目的生态周期，在数字人文领域使用关联数据技术能使研究人员以灵活的方式发布和使用数据。此外，也着重于通过对数据的重新组织构建，将其转化为能够支持领域研究的“智慧数据”，并形成全局知识网络，为社会公众、科研人员、科研机构等提供开源共享的智能知识服务 [9]。

根据以上文献收集后整理出的研究热点，下文将从数字人文领域数据资源建设、数字人文知识图谱构建技术、数字人文知识图谱平台智能应用三个方面进行详细讨论。

3 数字人文领域数据资源建设

数字人文领域资源建设需经过 3 个阶段，如图 1 所示。

第一阶段是进行数据集的构建，目的是实现资料的电子化，并以数据库等形式储存 [10]；第二个阶段是将结构化数据、半结构化数据以及非结构化数据转化成 RDF 结构化数据，实现语法层面的统一；最后一步则是通过本体融合和资源关联来实现关联不同数据源的资源，实现资源的分布式融合，进而实现语义层面的统一。

3.1 实现领域资源数字化

数据集的构建位于数字人文应用流程的基础阶段，GLAMs（Galleries, Libraries, Archives and Museums，艺术馆、图书馆、档案馆和博物馆）在数据积累方面有较大的优势，因此他们一般是数据集构建的主体机构，将纸质材料信息进行数字化并对其进行组织。数字人文数据

主要是文本形式，同时还有一些多源数据形式，例如图片、音频、视频、3D 等数据。针对不同的数据资源形式，也存在着不同的构建技术，下文将对不同的领域资源数字化过程进行分析。

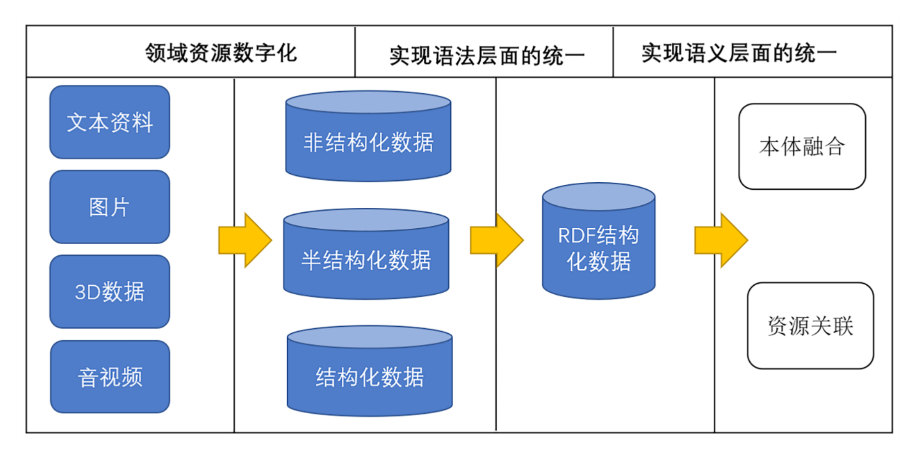


图 1 数字人文领域资源建设过程

（1）文本资料。文本资料包括地方古典文本资料、图书、笔迹、家谱资料等，这些文本资料需通过图像技术记录和保存原始文档的外观结构和内容，这一过程主要利用图像感光技术（Charge-Coupled Device，CCD）、图像传感技术（Complementary Metal Oxide Semiconductor，CMOS）等技术来对资源进行采集，这一过程需要与图像光学字符识别（ORC）结合使用，使图像转化为计算机可识别的 ASCII 码，再转化为文本资源，同时需要机器学习来实现识别任务。例如 M. Kestemont 等 [11] 着重研究中世纪拉丁手稿，通过卷积神经网络对手稿进行识别，并对自动分类的可行性进行了阐释。

（2）图片。图片包括地图、画作、壁画等，其电子化方法与文本资料类似，主要使用 OCR 与机器学习技术进行扫描与识别任务。如 S. A. Oliveira 等 [12] 着眼于 19 世纪初威尼托地区的拿破仑卡德斯地图，提出了第一个可以自动分割和解释 19 世纪初威尼托地区的拿破仑卡德斯地图的全自动系统，该系统使用机器视觉算法来提取出每个碎片的几何图形，并进一步对手写的标签进行分类、读取和解释。

（3）3D 数据。3D 数据有文物、器皿、雕塑等。3D 数据数字化是利用摄影、数字化扫描及编辑等最新的技术手段对信息进行数字化存储或重新构建三维数字模型，最后使用相关软件进行数字化还原 [13]。三维扫描技术，可以根据需求，记录文物最真实、最全面的形态特征。如今，3D 扫描技术越来越多地应用于文物保护领域。这种方法使文物的展示和检索更加数字化。同时，该技术的应用也更有利于文物研究、文物共享和文物传播。这一方面国外起步较早，有影响力的项目多，国内尽管起步晚，但也取得了不少有效的成果。比较著名的项目是斯坦福大学曾经开展的“米开朗基罗项目”，该项目针对世界著名的雕塑进行三维扫描，对其进行数字化保护。

（4）音视频。音视频数据包括访谈、纪录片等多媒体数据。对音视频进行数字化即是利用技术对其进行扫描、翻拍、转录，进而实现数字化。近年来，声像档案抢救性保护逐渐成为重点研究方向之一，与此同时，结合数字技术也逐渐成为一种必然趋势 [14]。要使音频档案与视频档案得到长久保存并被更多人利用，数字化是一种较为可行的方法 [15]。因此，在音视频数字化的过程中，对其进行修复是其中非常重要的一个环节，例如内蒙古自治区档案馆通过 COOL EDIT PRO2.1 与 ADOBE AUDITION

知识管理论坛，2022（1）:87-100
DOI: 10.13266/j.issn.2095-5472.2022.008

CC 等修复软件对音频文件进行数字化修复，首先将音量标准化提高，其次进行音量降噪处理，最后手工干预残存噪点；至于视频修复，则要坚持“最小干预”的修复原则，在“听清楚、看清楚”的基础之上，最大化保留音视频档案的原始凭证作用[16]。

3.2 实现资源语法层面的统一

随着科技发展，人工智能、智慧数据等不断进入人们的视野，各行各业对其研究也不断加深，正推动着数字人文发展从“互联”走向“智联”。人文学科的数据资源类型多样、来源多源、数据海量、环境异构，因此在该领域进行数据资源建设需要实现语法和语义层面的统一，由此来有效解决存在的诸如数据异构、实体消歧、关联共享等问题，实现数据的语义增强和价值提升。

对于结构化数据，通常采用 RDB2RDF 的方法进行转换，如使用 D2R 工具、R2RML 映射语言[17]等。EXCEL 和 CSV 文件也具有结构化数据的特点，可以使用 OpenRefine 来进行数据转换。半结构化数据是介于结构化数据和非结构化数据之间的一种数据，可以被看成是结构化数据的一种形式，并不符合关系型数据库的数据模型结构，但包含相关标记，可以用来分隔语义元素以及对记录和字段进行分层，因此它也被称为自描述的结构。我们可以使用 XML2RDF 或 JSON2RDF 等工具来实现非结构化数据向 RDF 结构数据的转换，这一过程被称为 RDFizer 实现。非结构化的文本数据需要结合自然语言处理（NLP）和命名实体识别（NER）技术，抽取出结构化数据，再进行 RDF 转换。而对于图像和音频视频文件的结构提取，主要先通过目标检测识别出资源实体，再进行转换。

3.3 实现资源语义层面的统一

结构化、半结构化和非结构化的数据资源统一转化成 RDF 结构的数据后，只是达成了语法层面的统一，为实现语义层面的统一，为实现资源的分布式融合，还需要将本地 RDF 数据集与对外开放的关联数据资源进行关联。

不同数据源资源之间的语义关联，通常通过本体融合和资源关联两步来完成：

（1）本体融合。目前本体融合的研究主要集中于寻找本体之间的映射，随着本体技术的发展，通过本体概念、实例及属性之间的语义匹配机制和映射方法，实现本体最小元素之间的相似对应关系，从而实现本体的最终融合[18]。目前国内外对本体融合的研究越来越多，也有许多成熟的本体融合系统，如 PROMPT、GLUE 等。AnchorPROMPT[19] 是由斯坦福大学开发的用来寻找本体之间映射的工具，该工具首先进行概念比较，然后利用本体结构判断可能相似的本体成分，但是对于复杂概念和关系的本体映射，AchorPROMPT 则无法处理。GLUE[20] 是基于实例的本体映射生成系统之一，利用机器学习技术，根据分类本体寻找本体间 1：1 的映射。M. Lamé 等[21]提出一种新的本体对齐框架，能够使文化遗产数据提供者生成定义良好且形式化良好的术语。

（2）资源关联。不同机构在将实体数据进行 RDF 结构化的过程中，往往会用各自机构的域名来定义资源的 URL 地址，这些资源之间需要进行关联操作。可以使用 LIMES、SILK、LDIF 等工具和框架来进行不同资源之间的自动化关联，主要原理是通过机器学习和字符相似度的一些算法来进行资源属性值的对比。

4 数字人文领域知识图谱关键构建技术

4.1 数字人文领域知识图谱构建框架

关联数据和广义知识图谱都是用节点和边来表示实体和关系，本文主要探讨如何用关联数据来解释广义知识图谱中的技术。关联数据表示的语义知识图谱中的实体必须以 RDF 命名，不同图谱之间具有标准的 SPQRQL 查询语言，因此可以解决知识表示和网络服务问题。数字人文领域知识图谱与通用知识图谱的构建方法存在诸多不同，尤其体现在本体构建、知识抽取、知识融合等构建技术中。本节将知识

图谱的构建技术和数字人文领域的知识特点相结合，在通用知识图谱的结构框架基础上，对数字人文领域的知识图谱构建框架进行归纳，如图 2 所示：

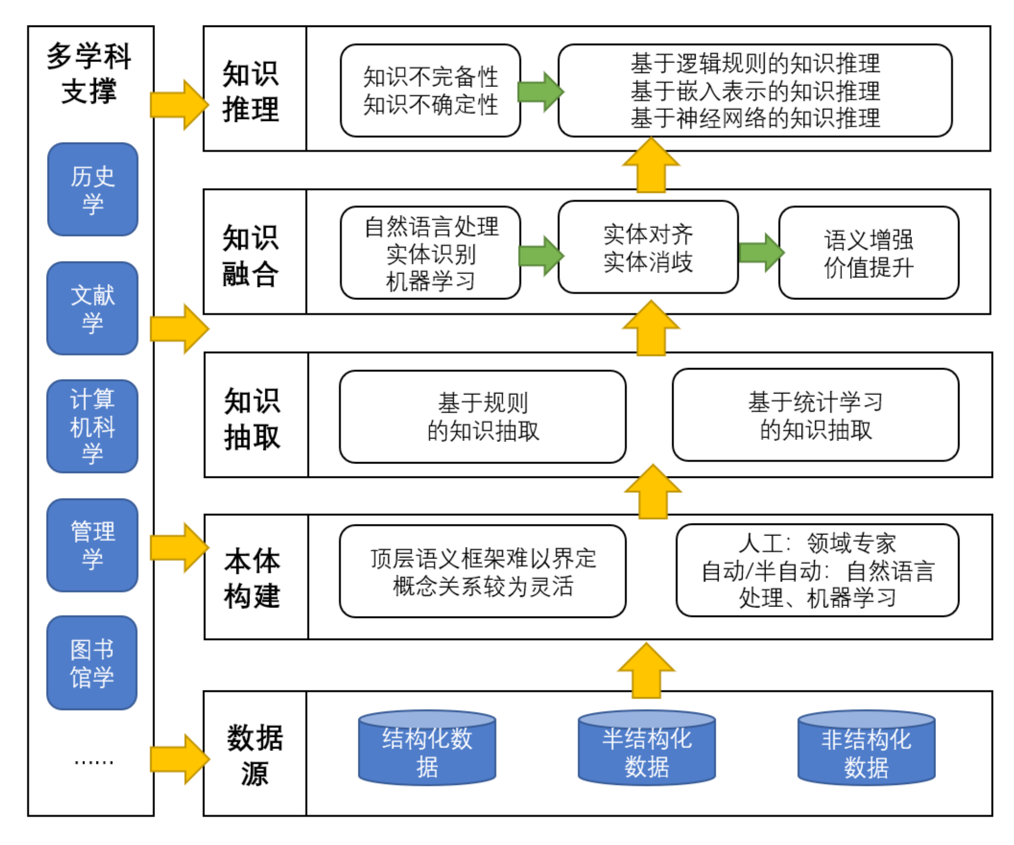


图 2 数字人文领域知识图谱构建框架

数字人文领域知识图谱构建框架主要包括多学科支撑基础、数字人文领域数据源、数字人文领域知识图谱构建。在多学科全方位的基础支撑下，基于海量、多元、异构的数字人文领域数据源进行本体构建、知识抽取、知识融合、知识推理，从而提供数字人文领域智慧数据产出。前文我们已经就数据资源的构建问题进行了分析，下文我们将针对数字人文领域知识图谱构建过程中的关键技术进行深入分析。

4.2 关键构建技术分析

4.2.1 本体构建

本体根据其描述的目标范围，可分为通用本体和领域本体。前者旨在建立可广泛应用于不同场景的本体知识，是对通用类知识的一种规范描述；后者则是对具体领域建立相对应的知识规范描述 [22]。

目前主流的本体构建方法分为人工构建和基于机器学习的自动化或半自动化构建两种。前者依靠领域专家的知识及经验，因此成本高且效率低下，与此同时，不同专家对同一事物的理解也不尽相同，因此人工构建的可拓展性较差。后者是指在已建立的本体语义框架下，结合自然语言处理、机器学习等技术从语料中自动抽取相关术语及属性关系，目前这种构建方法已经逐渐成为主流。

国外在领域本体的构建方法上的系统分析研究已经较为成熟，通过文献分析可知，国外典型的本体构建方法有 8 种，分别为：IDEF5 法、骨架法、TOVE 法、METHONTOLOGY 法、KACTUS 工程法、SENSUS 法、七步法以及循

知识管理论坛，2022（1）:87-100
DOI: 10.13266/j.issn.2095-5472.2022.008

环获得法 [23]。相较之下，国内起步较晚，技术相对落后，因此需要借鉴国外的构建方法，同时结合新的内容，形成新的观点方法。目前国内比较有代表性的本体构建方法主要有两种，分别是基于叙词表的构建方法和基于本体论工程法的半自动化与自动化构建方法 [24]。

近年来，一些学者构建了一些大型通用本体，如 DBPedia Ontology、YAGO 等。自然科学领域中大型实用化的领域本体发展迅速，因为其概念间的关系比较明确。目前比较有影响力的领域本体有 GeoNames Ontology、The Drug Ontology、UMLS SemNet、Gene Ontology 及 SNOMED 等 [25]。与自然科学领域不同，在顶层语义框架难以界定、概念关系较为灵活的人文社会科学领域中，大规模的实用化本体则较为少见 [26]。部分学者尝试开展对历史哲学等相关领域的本体构建研究，如国史本体、二十四史本体、哲学本体等 [27-28]；邓君等 [29] 针对档案领域构建了口述历史档案资源领域本体模型，有助于档案领域学者展开深层次研究；与此同时，在戏剧、民俗等领域，一些学者利用元数据、本体技术等进行信息资源描述和组织 [30]。

在语义环境下，领域本体的应用已成为一种必然，虽然国内目前的构建方法还不够完善，但自动化及半自动化的构建方法必将是未来的发展趋势。领域本体构建的进一步优化将着眼于以下几个方面：建立完善的评价机制，提高本体的重用性以及注重本体的共享性。同时，构建数字人文学科领域的大规模的实用化本体也将成为日后学者研究的重要方向之一。

4.2.2 知识抽取

随着自然语言处理技术的不断发展，数字人文领域内知识抽取的方法已经趋向于成熟，主要可以分为两个角度：基于规则的方法和基于统计学习的方法。

基于规则进行知识抽取的核心要点，就是关系规则的定义和规则两边的实体抽取，规则的精确度直接影响着所抽取知识的质量。在数字人文领域，基于规则的方法需要考虑词语之间的搭配关系和上下文语境。该方法具有准确率高、构建方法简单的优点。例如，刘悠然等 [31] 提出了一种基于规则的古汉语句型统计方法，该方法在标注高频字后，便能依据设定的约束规则对未标注字词进行标注并统计句型，从而简化古汉语研究过程中的人工统计工作。该统计方法在约束规则设置合理的情况下，对句型统计的正确率能够高于 95%。但是，该方法也同时具有诸多局限性。尤其是对于数字人文领域内的文本，规则的针对性比较强，也就代表着其泛化能力较弱。例如，谢明鸿等 [32] 提出了通过固定句式搭配规则来识别人物关系，但由于中文文本的表达方式十分多样，会出现预测结果和实际不一致的情况。如果需要获得更好的抽取效果，就要重新制定新的规则。因此，数字人文领域的研究者更倾向于采用基于统计机器学习的方法。

基于统计机器学习的方法在数字人文领域得到了越来越广泛的应用，相比于基于规则的方法，基于统计学习的方法不需要构建规则，一般都是自动地从训练语料中学习参数。例如，L. L. Liu 等 [33] 采用基于条件随机场的方法对用于历史研究的文学汉语命名实体的算法识别进行了研究。该方法在测试中的表现良好，从《地方志》中抽取出了大量人名和地名，用于丰富中国传记数据库（CBDB）。秦贺然等 [34] 利用 TextRank 模型对古汉语文本进行关键词抽取。通过实验，利用 TextRank 模型抽取了《春秋经传》中的关键词，准确度能达到 84%，这些关键词能够让数字人文领域的学者快速地了解到春秋时期的历史事件和春秋的时代面貌。并且，该模型的应用空间也十分广泛，不但能用于古汉语文本，而且也能应用于现代汉语，例如构建自动摘要系统。

综合来看，为了获取更丰富的数据以支持数字人文领域内知识图谱的构建，可以在抽取之前进行数据预处理，减少抽取时间，提高准确率。也可以将基于规则和基于统计的方法相结合，由于数字人文领域的实体和关系具有一

定的特征，可以通过人工少量标注之后，自动生成规则，同样也有利于提高领域内知识抽取的精度和效率。

4.2.3 知识融合

传统的知识融合问题主要涉及三方面，分别为知识融合框架、知识融合算法以及知识融合应用。知识融合算法可分为两类，分别是基于信息融合技术的知识融合算法和基于融合规则的知识融合算法，其中，大部分知识融合框架都是基于本体来构建的 [35]。知识融合算法基于信息融合技术和基于规则的知识融合算法。针对前者，很多研究都是借鉴信息融合算法，将其移植到知识融合中，构造针对知识融合的全新算法。基于 Bayes 方法、D-S 理论、蚁群优化算法的 3 种知识融合方法是融合决策处理的流行方法。周芳等 [36] 在知识管理领域中，通过融合处理，提高了结果可信度，并提升实现系统任务目标的能力。后者则是通过找寻信息之间的关联，用规则来进行知识表示。

而在数字人文领域，针对其特点，知识融合主要用于在不同来源实体间建立关联关系，将从多个分布式异构信息来源中发现的数据进行整合，同时进行识别和判断，消除可能存在的歧义、数据冗余和不确定性等问题，最终形成新的知识 [37]。知识融合可以有效解决在数字人文领域所存在的数据异构、实体消歧、关联共享等问题，实现数据的语义增强和价值提升。如陈涛等 [38] 在构建 CBDBLD（CBDB 关联数据平台）时，将转换的 RDF 数据与上海图书馆人名规范库、VIAF、DBPedia 等数据集进行关联，采用 SILK 或者 LIMES 框架进行关联；F. Frontini 等 [39] 提出了一种算法，来自动消除法国文学批评语料库中所被提及的歧义，其成功地将通用知识库（如 DBpedia）与特定领域的知识库结合在一起。

4.2.4 知识推理

知识推理是针对知识图谱中已有事实或关系的不完备性，挖掘或推断出未知或隐含的语义关系。一般而言，知识推理的对象可以为实体、关系和知识图谱的结构等。目前主要有基于逻辑规则的知识推理、基于嵌入表示的知识推理以及基于神经网络的知识推理三类方法。作为知识图谱的核心功能之一，知识推理为解决数字人文历史性所带来的知识的不完备和不确定提供了思路，但在当前的数字人文项目中还少有成熟应用。

基于路径规则的知识推理通过随机采样提取到的关系路径特征来提高计算效率，但是降低了知识图谱中信息的利用率；同时利用监督学习方法建立的关系推力模型很大程度上会受到训练数据的影响。对此，刘峤等 [40] 提出双向语义假设，对全局关系进行推理，结合局部模块进行加权合并，最终得到完整的逻辑规则推理算法。周莉娜 [41] 提出了面向本体构建的领域知识推理框架，通过 TPO4DK 模型，构造形式化的推理规则，对唐代诗人之间以及诗歌—诗人本体中的诗人流派属性、诗歌题材与主题属性进行知识推理，实现对唐诗文献学的版本证伪的应用。陆泉等 [42] 提出一种基于 OWL 语言的模糊本体表现模型，通过 SWRL 语言表示精确规则和模糊规则，构建面向知识发现的推理模型。该模型可以同时描述精确知识和模糊知识，简化了对模糊知识的表示和处理；同时，数字人文资源所蕴含的多源异构数据，特别是图像数据资源之间的语义关系和概念层次结构也推动领域内的知识推理，如周知等 [43] 参考 Eakins 图像语义层次模型和王晓光等人提出的数字图像语义描述层次模型 [44]，对图像资源的语义进行了多层描述，实现实体之间、概念之间的深度关联，满足知识推理的需要。

基于嵌入表示的知识推理技术优势同样明显。通过将图结构中隐含的关联信息映射到欧氏空间，使得原本难以发现的关联关系变得显而易见。因此，基于嵌入表示的推理是知识图谱推理技术的重要组成部分。基于神经网络的知识图谱推理，充分利用了神经网络对非线性复杂关系的建模能力，能够深入学习图谱结构特征和语义特征，实现对图谱缺失关系的有效

知识管理论坛，2022（1）:87-100
DOI: 10.13266/j.issn.2095-5472.2022.008

预测。一般地，应用于知识图谱推理的神经网络方法主要包括 CNN 方法、RNN 方法、图神经网络（Graph Neural Networks，GNN）方法、DRL 方法等[45]。

5 数字人文领域知识图谱平台智能应用

5.1 相关平台项目概述

在信息技术飞速发展的背景下，信息获取、存储和传播的方式都产生了巨大变革，数据成为数字人文研究的基础与核心之一，因此，数字人文学者对于领域内研究资料的处理方式也产生了翻天覆地的变化。在传统的人文研究中，学者往往注重数据的收集与整理。但由于数字化技术的欠缺以及原始资料本身的质量问题，学者整理出来的数据经常是不完整、碎片化的。在数字化技术得到深入发展之后，人文领域的数据虽有了较为快捷与全面的收集，但仍然是杂乱的，并不利于领域内学者的研究。随着数字人文领域知识图谱规模的逐渐扩大，传统的关系型数据库无法有效管理其中的数据。该领域学者的研究往往需要多个数据集的交叉查询，例如图像、文字、音频等数据之间都存在一定的关联，发掘这些联系有助于人文研究的推进。因此，目前的研究一般采取关联数据技术（即语义知识图谱）来实现数字人文领域的数据管理。陈涛等[46]将关联数据技术与广义知识图谱进行了对比后指出，关联数据侧重于知识的发布与链接，与注重“挖掘”的广义知识图谱不同，关联数据技术更侧重于“推理”，即展示资源之间的关联关系。利用关联数据技术能够支持大规模、可重用的数字人文研究[47]，通过对数据的重新组织构建，将其转化为能够支持领域研究的“智慧数据”，并形成全局知识网络。表 1 列举出了国内外数字人文领域平台建设的几个典型代表。

表 1 数字人文关联数据平台实践

领域类型	平台名称	发起机构	主要任务
历史	中国历代人物传记资料库关联数据平台（CBDB）	北京大学中国古代史研究中心、哈佛大学费正清东亚研究中心	展现了人物之间的亲属及社会关系，形成特有的社会关系网络，实现人物之间隐性关系的挖掘与呈现
	盛宣怀档案知识库	上海图书馆、上海科学技术情报研究所	收集政治、经济、社会、军事、外交、金融、贸易、教育各方面私人档案，是研究中国近代史的第一手史料宝库
档案	欧洲数字图书馆 Europeana	欧盟委员会	整合欧洲具有代表性的文化遗产资源，提供一站式浏览与检索服务，实现欧洲数字文化资源传播与共享
	中国家谱知识服务平台	上海图书馆	基于大量数据并结合时间、空间，对姓氏、人物及人物间的相互关系进行全景式的可视化展示和统计分析
	威尼斯时光机	瑞士洛桑联邦理工学院、威尼斯大学	将海量的历史档案进行数字化、转录、建立索引和关联
艺术	敦煌壁画叙词表关联数据服务平台	武汉大学数字人文中心	为敦煌壁画数字资源的深度语义标注、语义检索、知识组织、信息关联与共享等提供一套受控词表
	Getty 数字博物馆	美国盖蒂艺术中心	发布文物数字化建设的描述元数据标准和著录规范，各种数据值标准和数据交换标准
	中国传统建筑数字研究工具项目	范德堡大学数字人文中心	开发由开放数据库网站、建筑群、个体结构和结构元素四个相互关联部分组成的数字研究工具
文学	中国数字方志库	北京市文津书店	根据现有行政区划整体排列，涵盖宋、元、明、清及民国时期的刻本、抄本、稿本等各种版本的方志
	威尔士报纸在线	威尔士亚伯国家图书馆	将报纸档案数字化，目前包含大约 420 000 份来自威尔士和与威尔士相关的数字化报纸

从中可以看出，数字人文关联数据平台所横跨的领域十分丰富，主要有历史学、档案学、艺术、文学等。其中，历史学是数字人文平台实践最多的领域之一，而其他相关领域也与历史学有着千丝万缕的联系，能够体现出当今世界各国对于历史文化资源保存与利用的重视程度。

5.2 平台特点分析

5.2.1 跨界合作突出

国内外先进的数字人文关联数据平台一个突出的特点就是跨界合作，这是数字人文的跨学科属性所要求的。合作方式主要可以分为以下两种：

一方面是国内外机构的广泛合作。S. Wong[48] 指出，数字人文学科的合作性是该领域的核心价值之一，采用合作的方法可以利用各种机构的优势和专业知识，从而产生深远影响。比如欧洲数字图书馆 Europeana，有超过 15 个国家的 200 多个文化机构为该数字图书馆的开放数据集提供了贡献，包括伦敦的大英图书馆、阿姆斯特丹的里杰克斯博物馆和巴黎的卢浮宫等著名机构以及欧洲其他地方较小的文化遗产组织和图书馆 [49]。此外，由北京大学中国古代史研究中心与哈佛大学费正清东亚研究中心合作开发的中国历代人物传记资料库项目（CBDB），同样是国内外研究中心合作建立资料库的经典实践，该平台能够展现历史人物之间的各类关系，并形成特有的社会关系网络，实现人物之间隐性关系的挖掘与呈现 [50]，在研究中国历史的同时，能够促进西方国家对中国传统文化的理解。

另一方面是校外机构与高校的合作。大多数字人文机构隶属于大学，以高校图书馆依托进行平台建设，由高校图书馆、档案馆提供数据资源和人才，企业、基金会提供资金等。比如伊利诺伊大学香槟分校人文、艺术和社会科学计算所与亚伯拉罕·林肯博物馆合作开发的林肯著作数据库，该数据库由伊利伊诺大学香槟分校主导开发，投入人才资源支持与后续平台管理和服务，亚伯拉罕·林肯博物馆提供相关历史资源，形成人才资源与历史资源的相互支撑 [51]。学术机构、图书馆、档案馆、博物馆以及企业、基金会等之间建立广泛的联系，再加上人文、社科、理工等多学科参与，有利于资源的整合与创新利用。

5.2.2 实践导向性强

首先，较多数字人文关联数据平台为包括人文学科在内的一系列学科提供服务，例如提供数字化成像、数字保存、元数据创建、数据策展与管理、GIS 和数字映射、数字出版等多种数字学术功能。例如，由德国的柏林洪堡大学图书馆信息学院、曼海姆大学、开放知识基金会等多个机构合作研发的欧洲数字手稿项目，该项目构建了 DM2E 数据集，提供元数据和链接以及展示、处理、整合数据的相关工具，以便数字人文研究者和想要了解欧洲历史文化的群众直接访问欧洲各地各种文化遗产机构的数字化内容 [52]。这也体现了数字人文关联数据平台服务于实践，服务于解决实际问题的特点。

其次，这些平台都较为注重成果对大众的呈现与宣传。例如敦煌壁画叙词表关联数据服务平台通过叙词表可视化，降低了叙词表的认知难度，实现了专业化叙词表向适用于大众利用的过渡 [53]。上海图书馆研发的中国家谱知识服务平台 [54]，基于大量数据，采用时空结合对姓氏、人物及人物间的相互关系进行全景式的可视化展示和统计分析。由此可知，数字人文关联数据平台进行成果呈现一方面有助于数字人文研究的推广，提升数字人文学科影响力，另一方面有助于促进文化从现实世界向数字空间延伸拓展，丰富人类的数字文明内涵。

5.2.3 数据孤岛现象突出

数字人文关联数据平台数据资源的智慧性主要体现在及时性、可获取性以及可利用性 3 个方面。因此需要形成动态的、开放关联的数据资源，不断丰富其内容与形式。近年来，国内外对于数字人文关联数据平台建设越来越重视。但与此同时，新的隐患也在形成。王晓光

知识管理论坛，2022（1）:87-100
DOI: 10.13266/j.issn.2095-5472.2022.008

提出了数字人文研究中的数据失秩现象，尤其在中国大陆，这种现象更为严重，他指出：数字资源建设的主体走向多元化，图书馆、博物馆、档案馆等相关研究机构都投入了相当多的资金与人力支持，却导致了无数个更大的“数据孤岛”出现，比纸质文献时代更严重[55]。这种现象淡化了领域学者为平台建设所付出的相关努力，甚至可能给人留下一种数字人文研究的生命周期很短暂的印象。

纵观形成数据孤岛现象的原因，首先是随着研究的开展，资源数据量与研究资料的范围也在拓展。除了传统的文献资源以外，其他实物、图像、音视频等资料都会被列入数字人文学者的研究范围内。数字人文领域基础资料种类的繁杂容易造成相关研究的彼此孤立。其次，较多平台管理者倾向于将重点放在规划和启动新项目上，从而容易忽略对旧项目的后续管理、维护[56]。随着时间的推移，原有的数据资源格式可能会与现有的技术存在不相兼容的情况，旧的数据资源将无法与新的目标用户需求匹配。若不能及时更新现有的技术方法及操作环境，反而一味开展新项目，平台资源便很难保持鲜活。如何改善数据孤岛现象，实现对数字人文智慧数据资源的统一表示，已经成为数字人文智慧化知识服务平台发展道路上的重要议题。

6 数字人文领域知识图谱研究的未来趋势

综合近年来的数字人文领域知识图谱的研究成果，结合目前数字化技术的智慧化趋势，我们可以观察到如下发展趋势：

（1）多元数据集成。数据的长期保存是数字人文领域知识图谱平台非常重要的基础职能之一。与其他领域相比，数字人文领域中的数据相对来说比较特殊，包含了语言、文献、绘画、音乐等多种形式，它们的维度超越了可被物理上测量的范围，更加依赖于语义和语法[57]。对数字人文领域的研究离不开人文文献资料的数字化，庞大的数据资源在数字人文领域具有非凡的价值，而如何处理好这些数据，将其转换为机器可理解、可处理的资源至关重要。而数字人文研究只使用以往的数据资源是远远不够的，还需要大量鲜活的、正在被创造出来的数据。因此，可以利用社会性网络和开放存取的信息作为信息来源，将跨地域、跨学科、跨国别的联系变得更加紧密，在经过深度语义标注、结构化、形式化和可视化处理后，将数据转变为高级形式的智慧数据，并推进到更细化的分支领域。

（2）多模态知识融合。早期数字人文领域的多模态知识融合更多地针对不同知识源的各类知识，强调知识来源的多样性。未来，多模态知识融合将进一步突破传统的时间和空间限制，对于不同知识源的多样化特征进行涵盖与扩展，依托知识图谱智能平台的数据整合能力，打通文本、影像、实体（人物、地点、年代、地域、事件）等多维度语义资源，为体系化、语义化、系统化的数字人文资源整理、研究提供能力支撑。此外，对于同一知识源的不同解读也构成了数字人文资源的不同维度与层次，从而能够更好地满足数字人文领域研究中深层次的信息需求，并实现大数据环境下智能知识服务的不断创新。

（3）多学科交叉应用。数字人文领域关联数据平台构建的创新性研究应用于多种学科领域，有助于形成相互补充、相互验证的有机整体成果，能够将不同学科之间的距离缩小，促进学科的融合。一方面，学科的专业化程度不断提高，内部发展逐渐精细化，能够更具体、更深入地涵盖数字人文领域内容；另一方面，学科交融产生新的学科，如数字艺术、数字史学等。梁晨等[58]指出，数字技术或数据库平台还可以是微观信息的加速器或对撞机，并在数据的交叉和对撞过程中呈现出各种特征、趋势和规律。这些变化都在逐渐要求领域内研究人员不断突破不同专业之间的界限，为数字人文研究带来新的独有的研究范式，进一步推动交叉学科的稳固发展。

7 结论

从构建到为数字人文研究提供基础设施支持，数字人文领域的知识图谱研究经历了不断的发展与变革，以适应“数智时代”传统文献资源向智慧数据资源的转型。目前，数字人文领域知识图谱已经能够较好地提供知识发现和推理功能，支持多种类型的数字人文资源描述与融合，并能够满足文化的长期保存和共建共享的需求。本文以数字人文领域国内外会议、期刊发表的相关文献为研究对象，对数字人文领域的数据资源建设、知识图谱构建、智能服务平台 3 个方面进行调研，认识到数字人文领域知识图谱研究能够为该领域资源的数字化建设制定统一规范的方法参考，并为数字人文研究提供基础设施，更好地实现智慧数据资源的转型与升级。在这个过程中，新的机遇、新的挑战都在不断发生，而知识图谱作为人工智能时代一种先进的知识组织方式，能够充分发挥其知识融合中介的作用，为“数智时代”的发展提供源源不断的动力，并为我国未来的数字人文发展道路提供指引与方向。

参考文献：

[1] 李启虎 , 尹力 , 张全 . 信息时代的人文计算 [J]. 科学 , 2015, 67(1):35-39, 4.

[2] 刘炜 , 叶鹰 . 数字人文的技术体系与理论结构探讨 [J]. 中国图书馆学报 , 2017, 43(5):32-41.

[3] HASLHOFER B, ISAAC A, SIMON R. Knowledge graphs in the libraries and digital humanities domain[J]. arXiv preprint, 2018, arXiv:1803.03198.

[4] 周莉娜 , 洪亮 , 高子阳 . 唐诗知识图谱的构建及其智能知识服务设计 [J]. 图书情报工作 , 2019, 63(2):24-33.

[5] KAPLAN F. A map for big data research in digital humanities[J]. Frontiers in digital humanities, 2015, 2(1): 1-7.

[6] 董政娥 , 陈惠兰 . 数字人文资源调查与发展对策探讨 [J]. 情报资料工作 , 2015(5):103-109.

[7] 陈涛 , 刘炜 , 朱庆华 . 中文百科概念术语服务平台 SinoPedia 的构建研究 [J]. 中国图书馆学报 , 2018, 44(4):4-18.

[8] HOEKSTRA R, MEROÑO-PEÑUELA A, DENTLER K, et al. An ecosystem for linked humanities data[C]// European semantic Web conference. Cham: Springer, 2016: 425-440.

[9] Zeng M L. Smart data for digital humanities[J]. Journal of data and information science, 2017, 2(1): 1-12.

[10] 王军 , 张力元 . 国际数字人文进展研究 [J]. 数字人文 , 2020(1):1-23.

[11] KESTEMONT M, STUTZMANN D. Script identification in medieval Latin manuscripts using convolutional neural networks[C]// Premiere annual conference of the International Alliance of Digital Humanities Organizations. Montreal: McGill University, 2017.

[12] OLIVEIRA S A, KAPLAN F, DI LENARDO I. Machine vision algorithms on cadaster plans[C]// Premiere annual conference of the International Alliance of Digital Humanities Organizations. Montreal: McGill University, 2017.

[13] 张辉 , 王冬梅 . 基于三维扫描技术的唐陵雕塑数字化保护研究 [J]. 艺术与设计 (理论), 2016, 2(4):91-93.

[14] 刘江霞 . 模拟音视频档案数字化质量控制研究 [J]. 档案学研究 , 2018(1):101-106.

[15] 钱万里 . 传统声像档案的数字化处理 [J]. 档案与建设 , 2007(8):22-24.

[16] 罗永俊 , 毕晓然 , 郝阳 . 内蒙古民族文化珍贵音像档案抢救技术研究 [J]. 黑龙江档案 , 2020(5):43-45.

[17] R2RML: RDB to RDF Mapping Language [EB/OL]. [2021-07-23]. https: //www.w3.org/2001/sw/rdb2rdf/r2rml/.

[18] 熊顺 , 刘平芝 , 苏宗义 , 等 . 基于语义匹配映射的地理信息本体融合方法研究 [J]. 测绘科学与工程 , 2017 (1): 51-58.

[19] NOY N F, MUSEN M A. The PROMPT suite: interactive tools for ontology merging and mapping[J]. International journal of human-computer studies, 2003, 59(6): 983-1024.

[20] DOAN A H, MADHAVAN J, DHAMANKAR R, et al. Learning to match ontologies on the semantic Web[J]. The VLDB journal, 2003, 12(4): 303-319.

[21] LAMÉ M, PITTET P, PONCHIO F, et al. Heterotoki: non-structured and heterogeneous terminology alignment for digital humanities data producers[C]//Open data and ontologies for cultural heritage. Rome: Antonella Poggi, 2019.

[22] 任飞亮 , 沈继坤 , 孙宾宾 , 等 . 从文本中构建领域本体技术综述 [J]. 计算机学报 , 2019, 42(3):654-676.

知识管理论坛，2022（1）:87-100
DOI: 10.13266/j.issn.2095-5472.2022.008

[23] 尚新丽 . 国外本体构建方法比较分析 [J]. 图书情报工作 , 2012, 56(4):116-119.
[24] 岳丽欣 , 刘文云 . 国内外领域本体构建方法的比较研究 [J]. 情报理论与实践 , 2016, 39(8):119-125.
[25] WIMALASURIYA D C, DOU D. Ontology-based information extraction: an introduction and a survey of current approaches[J]. Journal of information science, 2010, 36(3): 306-323.
[26] 何琳 , 陈雅玲 , 孙珂迪 . 面向先秦典籍的知识本体构建技术研究 [J]. 图书情报工作 , 2020, 64(7):13-19.
[27] 王颖 , 张智雄 , 孙辉 , 等 . 国史知识的语义揭示与组织方法研究 [J]. 中国图书馆学报 , 2015, 41(4):55-64.
[28] THAKKER D, KARANASIOS S, BLANCHARD E, et al. Ontology for cultural variations in interpersonal communication: building on theoretical models and crowdsourced knowledge[J]. Journal of the Association for Information Science and Technology, 2017, 68(6): 1411-1428.
[29] 邓君 , 王阮 . 口述历史档案资源知识组织与关联分析 [J]. 情报资料工作 , 2021, 42(5):58-67.
[30] 周耀林 , 赵跃 , 孙晶琼 . 非物质文化遗产信息资源组织与检索研究路径——基于本体方法的考察与设计 [J]. 情报杂志 , 2017, 36(8):166-174.
[31] 刘悠然 , 龙丹 . 一种基于规则的上古汉语句型统计方法的设计与实现 [C]// 澳门大学人文学院、中国中文信息学会、澳门语言学会 . 第十五届汉语词汇语义学国际研讨会论文集 . 北京 : 外语教学与研究出版社 , 2014:428-433.
[32] 谢明鸿 , 冉强 , 王红斌 . 基于同义词林和规则的中文人物关系抽取方法 [J/OL]. 计算机工程与科学 , 2021, 43(9):1660-1667.
[33] LIU C L, HUANG C K, WANG H, et al. Mining local gazetteers of literary Chinese with CRF and pattern based methods for biographical information in Chinese history[C]//Proceedings of 2015 IEEE international conference on big data (Big Data), Santa Clark, 2015: 1629-1638.
[34] 秦贺然 , 王东波 . 数字人文下的先秦古汉语关键词抽取应用——以《春秋经传》为例 [J]. 图书馆杂志 , 2020, 39(11):97-105.
[35] 唐晓波 , 朱娟 . 大数据环境下知识融合的关键问题研究综述 [J]. 图书馆杂志 , 2017, 36(7):10-16.
[36] 周芳 , 刘玉战 , 韩立岩 . 基于模糊集理论的知识融合方法研究 [J]. 北京理工大学学报 (社会科学版), 2013, 15(3):67-73.
[37] 高劲松 , 梁艳琪 . 关联数据环境下知识融合模型研究 [J]. 情报科学 , 2016, 34(2):50-54.
[38] 陈涛 , 刘炜 , 单蓉蓉 , 等 . 知识图谱在数字人文中的应用研究 [J]. 中国图书馆学报 , 2019, 45(6):34-49.
[39] FRONTINI F, BRANDO C, GANASCIA J G. Semantic Web based named entity linking for digital humanities and heritage texts[C]// Proceedings of first international workshop semantic Web for scientific heritage at the 12th ESWC 2015 Conference. Portorož: Fabien Gandon, 2015:77-88.
[40] 刘峤 , 韩明皓 , 江浏祎 , 等 . 基于双层随机游走的关系推理算法 [J]. 计算机学报 , 2017, 40(6):1275-1290.
[41] 周莉娜 . 面向领域知识服务的唐诗本体构建与智能应用研究 [D]. 武汉 : 武汉大学 , 2020.
[42] 陆泉 , 刘婷 , 张良韬 , 等 . 面向知识发现的模糊本体融合与推理模型研究 [J]. 情报学报 , 2021, 40(4):333-344.
[43] 周知 , 蒋琳 . 数字人文图像资源知识组织模型构建研究 [J]. 图书馆学研究 , 2021(8):66-72, 65.
[44] 王晓光 , 江彦彧 , 张璐 . 敦煌壁画图像语义描述层次模型实证研究 [J]. 图书情报工作 , 2015, 59(19):122-129.
[45] 田玲 , 张谨川 , 张晋豪 , 等 . 知识图谱综述——表示、构建、推理与知识超图理论 [J]. 计算机应用 , 2021, 41(8):2161-2186.
[46] 陈涛 , 刘炜 , 单蓉蓉 , 等 . 知识图谱在数字人文中的应用研究 [J]. 中国图书馆学报 , 2019, 45(6):34-49.
[47] HOEKSTRA R, MERONO-PENUELA A, DENTLER K, et al. An ecosystem for linked humanities data[C]// Proceedings of European semantic Web conference. Cham: Springer, 2016: 425-440.
[48] SHUN HAN REBEKAH W. Digital humanities: what can libraries offer?[J]. Libraries and the academy, 2016, 16(4):669- 690.
[49] ISAAC A, HASLHOFER B. Europeana linked open data–data.europeana.eu[J]. Semantic Web, 2013, 4(3): 291-297.
[50] TSUI L H, WANG H. Harvesting big biographical data for Chinese history: the China Biographical Database (CBDB) [J]. Journal of Chinese history, 2020, 4(2): 505-511.
[51] Institute for Computing in Humanities, Arts, and Social Sciences[EB/OL].[2021-12-15].http://chass.illinois.edu/.
[52] BAIERER K, DRÖGE E, ECKERT K, et al. DM2E: a linked data source of digitised manuscripts for the digital humanities[J]. Semantic Web, 2017, 8(5): 733-745.
[53] 王晓光 , 侯西龙 , 程航航 , 等 . 敦煌壁画叙词表构建与关联数据发布 [J]. 中国图书馆学报 , 2020, 46(4):69-84.
[54] 夏翠娟 , 刘炜 , 陈涛 , 等 . 家谱关联数据服务平台的开

发实践 [J]. 中国图书馆学报 , 2016, 42(3):27-38.
[55] 王晓光 . 数字人文与智慧数据 [J]. 上海高校图书情报工作研究 , 2018, 28(2):25, 24.
[56] REED A. Managing an established digital humanities project: principles and practices from the twentieth year of the William Blake archive[J]. Virginia Tech, 2014, 8(1):1-17.
[57] SCHÖCH C. Big? smart? clean? messy? data in the humanities[J]. Journal of digital humanities, 2013, 2(3): 2-13.
[58] 梁晨 , 李中清 . 从微观数据到宏观历史：作为桥梁的数字史学 [J]. 中国社会科学评价 , 2021(2):84-92, 159.

作者贡献说明：
朱丽雅：参与框架制定，收集整理资料，撰写并修改论文；
张　珺：收集整理资料，撰写并修改论文；
洪　亮：提出论文主题和研究框架，指导论文写作；
罗绍辉：提出论文部分章节的写作思路；
兰　度：提出论文部分章节的写作思路。

Knowledge Graph in the Field of Digital Humanities: Research Progress and Future Trends

Zhu Liya[1] Zhang Jun[1] Hong Liang[1] Luo Shaohui[2] Lan Du[2]
[1]School of Information Management, Wuhan University, Wuhan 430072
[2]Surveying, Mapping and Geographic Information Institute of Nanning, Nanning 530022

Abstract: [Purpose/significance] This paper conducts a systematic review of the knowledge graph research in the field of digital humanities, aiming to provide possible future research directions and open research topics. **[Method/process]** By taking relevant paper published in domestic and foreign conferences and journals as the research objects and using the comprehensive induction method, the theoretical and practical development of the knowledge graph in the field of digital humanities was systematically combed. Then it explained the related concepts of the knowledge graph in the field of digital humanities. And according to the current research hot spots, this paper revealed its research trends from three aspects of the data resource construction, key construction technologies and intelligent application platforms. Finally, it showed the prospects for future research trends. **[Result/conclusion]** This paper summarized the future trends of the knowledge graph research in the field of digital humanities. In the future, it will show the development trends of multi-source data integration, multi-modal knowledge fusion and multi-disciplinary cross-application.

Keywords: digital humanities　knowledge graph　smart data; data resource construction　semantic mining